

---

# Probabilistic Programming for Malware Analysis

---

**Brian Ruttenberg**

Charles River Analytics  
625 Mt. Auburn St  
Cambridge, MA, 02138  
bruttenberg@cra.com

**Lee Kellogg**

Charles River Analytics  
625 Mt. Auburn St  
Cambridge, MA, 02138  
lkellogg@cra.com

**Avi Pfeffer**

Charles River Analytics  
625 Mt. Auburn St  
Cambridge, MA, 02138  
apfeffer@cra.com

## 1 Introduction

Many malware authors borrow source code from other authors when creating new malware, or will take an existing piece of malware and modify it for their needs. As a result, malware within a family of malware (i.e., malware that is closely related in function and structure) often exhibit strong parent-child relationships. Determining the nature of these relationships within a family of malware can be a powerful tool for cyber-defense.

Generating the *lineage* of a family of malware is thus an important task but is difficult to perform manually due to the sheer volume of malware, intentional obfuscation by malware authors, and the many features and subtleties that must be examined to determine parent-child relationships. In this work, we describe a novel method for generating the lineage of a large family of obfuscated malware. We formulate the problem as a generative probabilistic model and develop a probabilistic programming (PP) algorithm to learn and infer the temporal and structural organization of a family's lineage. We demonstrated the accuracy and validity of our approach on synthetic data and real lab-generated malware. This work has significant implications in the cyber-defense community, and presents a problem that would be difficult to solve without the benefits of PP.

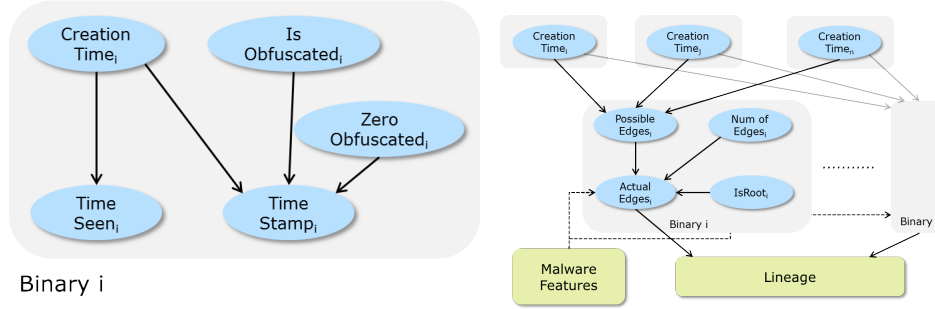
## 2 Lineage as a Probabilistic Model

The lineage of a set of malware binaries is a directed graph, where the nodes are the set of binaries in the family, and an edge from binary  $A$  to  $B$  implies that binary  $B$  evolved partly from  $A$  (and by implication, was created at a later time). The lineage can have multiple roots and binaries that contain multiple parents. By the definition of lineages, it is essential to know the order in which binaries were created. Without this information, it would be difficult to determine the inheritance direction of any parent-child relationships. As such, the lineage of a set of binaries (i.e., the graph) is conditioned upon the creation times of each of the binaries. We initially represent this simple relationship as high-level probabilistic model.

The *Lineage* variable represents a distribution over the possible lineages that can be constructed from a set of binaries, conditioned upon the *Creation Times* and the *Malware Features*. The *Creation Times* represents a distribution over the creation times of each binary and *Malware Features* is a deterministic variable that constrains the lineage generation based on binary similarity. The more features that malware binaries share, the more likely they are connected in the lineage, but the actual parent-child assignment of the two nodes depends upon the given creation times. Lineage on a set of malware  $\mathbb{M}$  is then defined as

$$Lineage_{\mathbb{M}} = \underset{Lineage_{\mathbb{M},i}}{\operatorname{argmax}} P(Lineage_{\mathbb{M},i} | Features_{\mathbb{M}}, Times_{\mathbb{M}}) \quad (1)$$

Computing Eqn. 1 on a set of malware is often difficult because the compiler time stamp is often purposely obfuscated by the author, so the creation times must be inferred using any available information, either from within a binary or using external information. Fortunately, we can also use the date that malware was first encountered in the wild as additional evidence.



(a) Plate for the model of each binary's creation time to infer a (b) Lineage model given distributions of the creation times of lineage-independent distribution of its creation time using time each malware. The malware features are used as soft constraints on the inheritance relationships.

Figure 1: Probabilistic models of lineage

One of the key insights is that the lineage and creation times are joint processes that can inform each other; knowing the lineage can improve inference of the creation times, and vice-versa. As such, performing joint inference of these models can potentially produce better results than inferring the creation times first and conditioning the lineage on the most likely creation times.

## 2.1 Creation Time Model

The plate for the creation time model is shown in Fig. 1(a). For a set of  $N$  binaries, we instantiate  $N$  independent models. While the binary creation times are *not* truly independent, the dependence between the creation times of different binaries is enforced through a joint inference algorithm, detailed in subsequent sections.

Each probabilistic model contains five variables. First, there is a variable to represent the actual binary creation time. There is also a variable to represent the time the binary was first seen in the wild. There is also a variable to represent the time stamp of the binary (from the actual binary header). This variable depends upon the creation time, as well as two additional variables that represent any obfuscation by the malware author to hide the actual creation time; one variable determines if the time stamp is obfuscated, and the other represents how the time stamp is obfuscated (either empty or some random value). Evidence is posted to the time seen and time stamp variables and a distribution of each malware's creation time can be inferred. Note that the priors for the obfuscation variables and parameters for the conditional distributions can be learned.

### 2.1.1 Lineage Model

The model for lineage is shown in Fig. 1(b). For each malware binary, we create a set of variables that represent how the binary can be used in the lineage. First, there is a variable which represents the possible existence of edges between a binary  $i$  and all the other binaries. This variable is deterministic conditioned upon the creation times of all the binaries

There are also two variables that control the number of edges (i.e., parents) for each binary, as well as a variable that specifies whether a particular binary is a root in the lineage. Finally, there is a variable that represents a set of actual lineage edges of a binary  $i$  which depends upon the possible edges of the binary, the number of edges it has, and whether it is a root binary. By definition, the values of the actual edges variable for all binaries defines the lineage over the set of malware (i.e., it can be deterministically constructed from the edges and the creation times.)

The conditional probability distribution of the actual edges variable is constrained by the difference between the features of the binaries. That is, the higher similarity between two binaries, the more likely they are to have an edge between them. The similarity measure between binaries is based on binary similarity measures and we refer the reader to [3] for more details.

## 3 Inference Algorithm

As shown in Eqn. 1, the lineage is the maximal probability lineage given the creation times and the malware features. Since the creation times are unknown, we must infer both the lineage and the

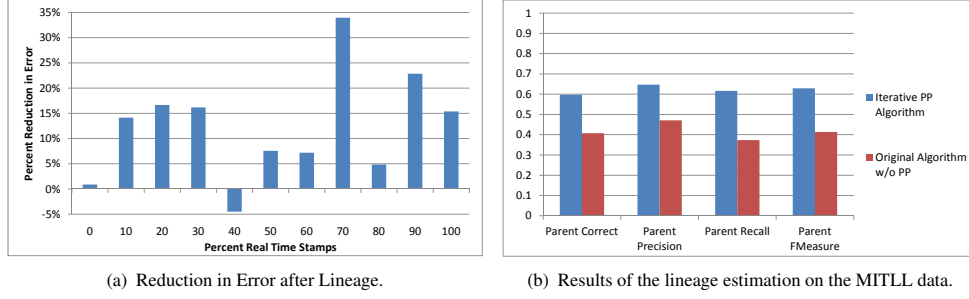


Figure 2: Testing results

creation times jointly. To accomplish this, we employed an iterative algorithm to jointly infer the most likely binary creation times and lineage, outlined as follows:

1. **Infer a distribution of the binary creation times.** Using the observable time stamp and time seen information, we infer a distribution of the creation times of each binary. This distribution is still conditioned upon the lineage; this process marginalizes out some of the information not needed to compute a lineage.
2. **Sample the creation times.** We take a sample from the creation time distributions of the malware binaries. This creates a fixed order of the binaries.
3. **Infer the most likely lineage.** We infer the most likely lineage of the malware binaries given the fixed creation times. That is, we compute the lineage described in Eqn. 1.
4. **Infer the most likely creation times.** The most likely creation times is defined as the set of creation times that maximizes the probability of the creation times conditioned on the features and the current lineage. Since we are conditioning on the previously computed lineage, we fix the inheritance between two binaries, but the direction of the edge can change depending on the inferred creation times.
5. **Repeat steps 3–4 until convergence.** We repeat the process until convergence.
6. **Repeat steps 2–5 until enough samples have been collected.** Since there is no guarantee that the maximization process converges on the global maximum, we restart the process to increase the likelihood that the global maximum is reached.

The algorithm is very similar to the expectation–maximization algorithm [2], but we must re-sample several initial malware creation times to ensure that we are finding a satisfactory maximum value. At the end of the algorithm, we select the lineage with the highest probability (from the multiple restarts of the algorithm) as the lineage of the set of malware.

## 4 Implementation and Testing

We implemented the model in Figaro [5], an open-source PP language, which provided several benefits on this problem. First, many of the model structures are repeated, so Figaro’s object-oriented nature facilitated easy creation and reuse of these structures. Second, Figaro is a natural environment to create the iterative maximization algorithm; Figaro’s inference engine can reason on any model encoded in the language, so it is easy apply successive maximization algorithms to the model. Finally, Figaro has several built in algorithms for learning and maximization, allowing us to focus on the model representation and not on the inference.

Inferring the creation times depends upon variable parameters which are generally unknown. For instance, we need to know the prior probability that a binary’s time stamp has been obfuscated. To determine these parameters, we learn them on a training set of data. Standard Bayesian machine learning methods can be used to learn the parameters, and we omit details for brevity.

Maximization in the iterative algorithm used Simulated Annealing to compute the most likely values for both the creation times and lineages. The Metropolis–Hastings [4] algorithm was used to infer the distribution of creation times for the binaries.

We tested the lineage method on both synthetic data and real malware generated by MIT Lincoln Lab (MITLL) as part of the DARPA Cyber Genome effort [1]. The synthetic data was primarily used to compute the accuracy of the creation time learning and estimation, and our results showed that our probabilistic model is an effective means to capture this information. More importantly, we also used the synthetic data to verify if the iterative algorithm can improve the creation time estimation (i.e., is there benefit to joint inference). In Fig. 2(a) we show the reduction in the creation time error (expected creation time compared to real creation time) after lineage is performed, where we varied the obfuscation of the malware time stamps from 0% to 100%. As can be seen, on all but one of tests, our estimate of the creation times improves after the iterative lineage algorithm has run, demonstrating the value of joint inference on this problem.

We tested the lineage estimation algorithm on 17 lineages generated by MITLL. Fig. 2(b) the accuracy of the lineage algorithm on four metrics. We show the results of lineage generation on the same 17 families using our original implementation that did not use PP or joint inference (it computed lineage using a single estimate of the creation time). As can be seen, our lineage estimation algorithm can reconstruct lineages of malware with fairly high accuracy, and is a significant improvement over the non-PP solution. These results demonstrate that our probabilistic formulation of the lineage problem is an effective solution to construct lineages in light of uncertain and obfuscated data.

## Acknowledgments

This work was supported by the Defense Advanced Research Projects Agency (DARPA) under US Air Force contracts FA8750-10-C-0171 and FA8750-14-C-0011, with thanks to Mr. Timothy Fraser. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. Approved for Public Release, Distribution Unlimited.

## References

- [1] DARPA. Cyber defense (cyber genome) program, 2010.
- [2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.
- [3] A. Lakhota, M. D. Preda, and R. Giacobazzi. Fast location of similar code fragments using semantic ‘juice’. In *Proceedings of the 2nd ACM SIGPLAN Program Protection and Reverse Engineering Workshop*, PPREW ’13, page 5:15:6, New York, NY, USA, 2013. ACM.
- [4] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.
- [5] A. Pfeffer. Creating and manipulating probabilistic programs with Figaro. In *2nd International Workshop on Statistical Relational AI*, 2012.